Project Progress Report
MassTrack: The MBTA Green Line
Tony Faller (af3370)

Link to Private Systems Project Doc: [Privacy Attacks - Public Data] MassTrack
Link to MBTA Dataset:
https://mbta-massdot.opendata.arcgis.com/datasets/5f71a5c035fc4a4dad1b7fa73ba27ef8/about

Overview
This project is an extension of a project which I am concurrently developing in the Private Systems course. Please see the above document and prior proposal submissions for further detail on the original project. In summary, in Private Systems I am analyzing BlueBike data for privacy vulnerabilities; in this course, I am analyzing MBTA data for similar privacy vulnerabilities. Note that as of this writing, there do not appear to be any overlaps between the two projects. However, if an overlap occurs, I will mark it clearly.

I would like to determine whether or not certain combinations of fields in the MBTA data constitute pseudo-identifiers – that is, a sort of "key" which can be used to uniquely identify an entry in the dataset. The existence of pseudo-identifiers implies that the data is vulnerable to attack by an adversary, given certain side information; for example, an adversary A who sees person B enter a train at station S and notes the time T could use the MBTA data to later determine where B went.

To assess whether or not certain combinations constitute pseudo-identifiers, I plan to determine the number of unique rows grouped by each combination. A higher percentage of unique rows strongly suggests that particular combination is indeed a pseudo-identifier, meaning that if the adversary knows this information independently, the adversary will be able to uniquely identify that particular entry in the dataset and learn new information such as the next station of the ride.

To demonstrate the risks associated with pseudo-identifiers, I plan to develop a system known as MassTrack. Given a real start station, start time, and time spent in transit, MassTrack will output a list of possible end stations for that particular ride. This system will be measured for its execution time and its accuracy (i.e. is the true end station within the output list, and how many stations are output). A lower execution time coupled with a higher degree of accuracy clearly demonstrates that this dataset is vulnerable to attack.

For simplicity, this project will be limited to the MBTA's Green Line trains.

Research Questions

1. Does the combination of [*to_stop_arrival_datetime*, *to_stop_name*] constitute a pseudo-identifier?
2. Does the combination of [*to_stop_arrival_datetime*, *to_stop_name, direction*] constitute a pseudo-identifier?
3. How accurate is the MassTrack system?
4. How does time-in-transit affect the execution time of the MassTrack system?
5. How does time-in-transit affect the accuracy of the MassTrack system?

Value to User Community

This project is valuable to privacy researchers and regular train riders. In Private Systems, I've learned a common phrase "private data isn't" – which means that "private data" is either non-private, or it is obfuscated to the point of not being useful as data. This is further complicated by the notion of "side information," which grants the attacker knowledge outside the confines of the dataset.

If this data turns out to be vulnerable from a privacy standpoint, it is a real-world example of "private data isn't." There is nothing in the dataset itself which can be used to identify a single individual – indeed, this dataset is associated with trains, and not with humans. However, this data may yet still be leveraged for an attack against the privacy of an individual.

Anyone who rides the T regularly may also be interested in this project. For particularly concerned individuals, the results for which fields constitute a pseudo-identifier are valuable. If these individuals find themselves within these groups (i.e. they typically leave a certain station around a particular time), then they are a more vulnerable population.

This project will be developed in Python and uploaded to a GitHub repository. The dataset is available on Zenodo. The links are as follows:

- https://github.com/amfaller/MassTrack-MBTA
- https://zenodo.org/records/15121997

A stretch goal of this project is to produce a deliverable in the form of some sort of application. This application would have input fields for start time, start station, and time-in-transit. When the user enters these, the application will display the list of output stations.

Demo

I plan to have slides for introduction/overview and results. If the above stretch goal is met, then the demo will also include some examples of tracking. I plan to submit this as an asynchronous video.

Delivery

A GitHub repository and Zenodo link (see above).